DOCUMENT RESUME

ED 357 080 TM 019 859

AUTHOR Lofald, Daniel R.; Pajares, M. Frank
TITLE The Effect of Embedded Questions on Readers' Calibration of Test Readiness.
PUB DATE Apr 93
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Beliefs; Cognitive Processes; Expository Writing; Higher Education; *Objective Tests; Performance; *Reader Response; Reader Text Relationship; *Readiness; *Self Evaluation (Individuals); Test Coaching; Test Wiseness; *Undergraduate Students
IDENTIFIERS Calibration; *Embedded Items; Questions; Subjective Evaluation; *Test Readiness

ABSTRACT

Whether questions embedded in expository text could improve the correspondence between adult readers' subjective assessments of test readiness and their objective test performance (prediction calibration) was studied with 168 undergraduates. In order to minimize the confounding effects of prior knowledge, the subjects were asked to read a text based on a make-believe solar system. This experiment was prepared as a two-factor design, embedded questions (yes/no) and text reinspection (yes/no). Because subjects were not cued to process the questions in any fashion, effects discovered were learner-produced, rather than investigator-induced. The purpose of the lookback factor was to separate the effects of embedded questions on perceptions of cognitive readiness when combined with re-study decisions from the effects of embedded questions when re-study was prohibited. Embedded questions had the effect of bringing subjective beliefs regarding test readiness into better calibration with objective test preparedness, and may thus be used to change the passive and dysfunctional relationship many readers have with the text. One figure illustrates the discussion, and six tables present study data. (Author/SLD)

THE EFFECT OF EMBEDDED QUESTIONS ON READERS' CALIBRATION OF TEST READINESS

Daniel R. Lofald

Education Division, Campbellsville College

M. Frank Pajares

Department of Foundations, University of Florida

Date of Submission: 4/3/93

Correspondence concerning this article should be submitted to

Dr. Daniel R. Lofald
Education Division
Campbellsville College
200 West College Street
Campbellsville, KY 42718

Tel: 502-789-5252
Fax: 502-789-5020

Running Head: CALIBRATION

Daniel Lofald is an assistant professor of educational psychology, learning and cognition and child development at Campbellsville College in Kentucky.

Frank Pajares is an adjunct professor at the University of Florida, Gainesville. His areas of specialization are educational psychology, child development, and teacher education.

THE EFFECT OF EMBEDDED QUESTIONS ON READERS' CALIBRATION OF TEST READINESS

Abstract

Perceptions of comprehension and cognitive readiness are salient features of academic reading and studying and the principal determinants of the learning strategies readers employ and cognitive resources they expend. Undetected cognitive failure during reading is a problem well-documented with young readers, and researchers have recently established that even adult, skilled readers are often not proficient at monitoring their cognition. This experiment examined whether questions embedded in expository text could improve the correspondence between adult readers' subjective assessments of test readiness and their objective test performance (prediction calibration). In order to minimize the confounding effects of prior knowledge, subjects were asked to read a text based on a make-believe solar system. This experiment was prepared as a two-factor design, embedded questions (yes/no) and text reinspection (yes/no). Because subjects were not cued to process the questions in any fashion, effects discovered were learner-produced rather than investigator-induced. The purpose of the lookback factor was to separate the effects of embedded questions on perceptions of cognitive readiness when combined with re-study decisions from the effects of embedded questions when re-study was prohibited. Subjects were 168 college undergraduates. Embedded questions had the effect of bringing subjective beliefs regarding test readiness in better calibration with objective test preparedness and may thus be used to change the passive and dysfunctional relationship many readers have with the text.

Submitted

THE EFFECT OF EMBEDDED QUESTIONS ON READERS' CALIBRATION OF TEST READINESS

Reading expository prose is one of the primary mechanisms through which students acquire knowledge in academic settings, with the principal motivation usually being to prepare for a test. Students read and study the text until they believe they have learned the material. When these judgments of test readiness are correlated with actual performance, the learner is said to be calibrated (see Glenberg, Sanocki, Epstein, & Morris, 1987; Lichtenstein, Fischoff, & Phillips, 1982).

Perceptions of comprehension and cognitive readiness are important features of academic reading and studying and the principal determinants of the learning strategies readers employ and the cognitive resources they expend. For example, if, during the course of reading, readers detect that their comprehension has failed, they are likely to engage in some form of remedial behavior. If comprehension failure goes undetected, the reader will not engage in strategic behavior and may stop reading before the material has been learned. Students who do not detect cognitive failure during reading are likely to do poorly on initial tests and progressively worse as the text and test incorporate earlier concepts.

Researchers have documented that undetected cognitive failure is a problem common to younger, less experienced readers (Baker & Brown, 1984a, 1984b; Brown, Armbruster, & Baker, 1986; Wagoner, 1983), but college students often perform poorly on examinations for which they feel adequately prepared. Although there are explanations for this beyond undetected cognitive failure, experimental findings suggest that college students often do not detect comprehension failure during reading relative to a given task

(Epstein, Glenberg, & Bradley, 1984; Glenberg & Epstein, 1987; Glenberg, Wilkinson, & Epstein, 1982; Maki & Berry, 1984; Pressley, Snyder, Levin, Murray, & Ghatala, 1987).

In most studies, readers are either asked to make predictions of performance before taking an examination (prediction calibration) or estimate performance at various points during the test itself (postdiction calibration). In the first instance, correlations between predictions and actual performance have been near zero (Glenberg & Epstein, 1985). Correlations between postdiction judgments and actual performance, however, have been much larger (Metcalfe, 1986). Some researchers have suggested that the test provides feedback that readers use to make more accurate judgments (Glenberg, et al., 1985, 1987; Pressley et al., 1987; Walczyk & Hall, 1989). Judgments before testing should be the focus of research, for only they influence decisions to use different study strategies and/or expend greater effort. This experiment explored instructional interventions to assist readers during reading, with the goal of bringing subjective judgments of test readiness in closer correspondence with an objective assessment of test performance.

In previous research, the challenge has been to externalize these complex mental events so they can be studied. Researchers (Baker & Brown, 1984a) have attempted to understand these mechanisms by trying to make a connection between what readers can declare about the workings of their memory with subsequent memory performance. Results using this approach have been disappointing. The central theoretical argument forwarded in this study is that this tactic is inappropriate for studying perceptions of cognitive readiness; readers may not be able to give accurate accounts of their memory processes because many of the cognitive events associated with memory monitoring reside within the executive processes. They are mental events not always available to conscious awareness and, therefore, not easily reportable. For these reasons, readers were asked to make comparative decisions about their perceptions of test readiness. This tactic captures the internal workings of metacognitive knowledge and executive processes while avoiding

the problems associated with other approaches--decisions regarding cognitive readiness reflect, in-part, a reader's declarative knowledge concerning readiness, but the procedure is not dependent upon the reader's ability to report on this knowledge while still capturing the outcomes of executive decision making.

The Role of Embedded Questions in Text

Questions embedded in text have been found effective in the learning of prose materials (Andre, 1987) and may affect cognitive monitoring by altering readers' perceptions of a reading task. Pressley et al. (1987) tested the effect of embedded questions on perceptions of cognitive readiness and found that embedded questions made estimates of prediction calibration statistically more accurate.

Researchers suggest that embedded questions improve assessments of test readiness in a number of ways. For example, the cognitive processes used to answer test questions may provide feedback that learners use to make more accurate judgments of their cognitive readiness (Glenberg, et al., 1987). Embedded questions may cause readers to evaluate their comprehension where they may not have done so on their own. If they cannot answer the embedded questions, they may readjust their learning strategies. Embedded questions may also serve as a benchmark by informing readers of the semantic level to which they should direct their reading. Although isolating and testing these effects can be difficult, each shares the outcome of altering readers' perceptions, a variable more easily measured. Researchers have manipulated task characteristics or orientating instructions and altered readers' ratings of perceived comprehension (Pratt, Liszcz, MacKenzie-Keating, & Manning, 1982; Shaughnessy, 1981) as well as their reported sense of cognitive readiness (Brown, Bransford, Ferrara, & Campione, 1982).

Our purpose was to investigate the effect of embedded questions on perceptions of cognitive readiness as well as on a number of other perceptions. When this question was

tested previously (Pressley et al., 1987), readers were given explicit instructions to answer the embedded questions. Most reading settings, however, are learner controlled and subject to the nuances of reader decision making. What has yet to be determined are the effects of embedded questions on calibration of comprehension when readers are not explicitly cued. The purpose of this experiment was to determine whether questions embedded in expository text improve the correspondence between adult readers' subjective assessments of test readiness and their objective test performance (prediction calibration).

Method

Subjects and Materials

Participants were recruited from the college of education of a large, public university in the South. Participation was voluntary and the final sample consisted of 168 students. Bias stemming from recruiting subjects from different classes under different incentive conditions was minimized through random assignment to experimental conditions. The experiment was conducted in a quiet, well-lit office in the College of Education. One of the authors, who was blind to group membership, supervised every session.

The experimental task for all groups was to read and study a text called Xenograde Science (Merrill, 1965), which was administered through an Apple Macintosh computer using the HyperTalk™ programming language. Although the experimental situation consisted of students reading text from a computer screen and recording their responses with a computer keyboard, no computer literacy was required and typing skills were reduced to depressing two keys on the keyboard. Xenograde Science (Merrill, 1965) describes the physics of a make-believe solar system. A contrived task was used for two reasons: (a) to minimize the effects of domain specific prior knowledge and (b) because novel material has been reported to produce more conscious, analytical processing than material more familiar to subjects (Hare & Smith, 1982).

The Xenograde task consisted of 20 frames of text totalling approximately 1700 words. There were 19 diagrams and 2 tables. In the two embedded questions conditions, after every fifth frame of text a frame containing the embedded questions was included. There were 18 distinct questions in these five frames, adding approximately 280 words to the task. The first embedded questions asked about the nomenclature of the Xenograde solar system; other questions asked readers to make predictions based on principles they should have learned as part of their reading. The text was presented one page at a time on the computer screen, and readers controlled page turning by pressing designated keys on the keyboard. A criterion test consisting of 17 multiple-choice questions was administered. All questions required subjects to apply the principles they had learned while reading the material.

Procedures

Subjects were first read a common script introducing them to the task and experimental objectives. They then picked a unique identification number from a box. Their attention was then directed to the first frame of the computerized textbook and they were instructed to proceed. On the seventh page of these instructions, subjects entered their identification number into the computer and were routed to the appropriate experimental condition.

After 23 minutes, or when participants had determined they had studied the material adequately (whatever condition occurred first), the computer routed them to the questionnaire that measured the independent variables. Subjects responded to the questions as they were presented, and their answers were recorded on a remote text file in the computer. After finishing the questionnaire, the computer routed them to the Xenograde test.

The decision to limit the amount of time subjects could read and study the text was motivated by the fact that traditional educational settings usually operate under the same

constraint--time allocated to learn a given quantity of material is limited (Chronbach & Snow, 1977).

Criterion Variables and Measurement Techniques.

By asking readers to make decisions about the need to reread, Pressley et al. (1987) created a measure based on the "perceived readiness for examination performance," (p. 222). or PREP, which they operationalized by asking readers to provide judgments of their perceived need to reread (would need to/would not need to) if criteria for passing the examination were 20, 40, 60, or 80%. Readers also indicated how confident they were in their decisions on a 5-point Likert scale from 1 (low confidence) to 5 (high confidence) to each of the 4 probes. For example, consider a hypothetical reader who scored 70% on the criterion examination and responded to the 4 probes as shown in Figure 1.

Insert Figure 1 about here

The PREP score for this reader would be the sum of (5%, +2, -2, +4), or 9. Using this procedure, PREP scores can range from between -20 and +20. Correspondence between these judgments and subsequent test performance forms the basis for a calibration score based directly on the perceived need for strategic remedial behavior. This experiment used the same PREP measure but altered many of the experimental conditions to determine if these findings would hold under conditions more similar to the academic reading situations normally encountered by adult learners.

Subjects were also asked to predict the score they would obtain on the test after studying the Xenograde text. This subjective assessment of cognitive readiness was compared to subsequent test performance. In this study, the discordance between the predicted and subsequent objective scores has been operationalized as prediction inaccuracy (PI) and is calculated by taking the difference between these two scores. Results from the

PREP instrument, PI measures, and the confidence scores readers assigned to their test responses served to determine if embedded questions improve calibration.

Research Questions and Data Analysis.

The research hypotheses were grouped into three conceptually distinct experimental units corresponding to the three dependent variables under investigation (PREP, PI, Criterion Test Performance). Each test of main effects was proceeded by an examination of the interaction hypothesis. If there was an interaction between the two factors (embedded questions and lookback), pairwise comparisons of cell means were conducted. If no interactions were found, a two-way analysis of variance (ANOVA) was conducted to test for main effects. The interaction parameter was tested in each model. The familywise error rate was controlled within this conceptual unit. Because effects were expected in each of the hypotheses outlined, a directional alternative was used in tests of significance.

The Xenograde experiment consisted of less than 23 minutes of reading wherein these perceptions could materialize. In addition, it is unlikely that the subjects had any interest in either learning the experimental materials or doing well on the examination. If facilitating effects of embedded questions could be discovered under these conditions, we would expect that similar or more robust effects might be found in more typical, high-stakes academic reading situations. The following research hypotheses were tested:

1. The mean PREP score for the embedded question groups will be significantly larger than the mean PREP score for the no-embedded question groups.

2. The mean PREP score for the lookback groups will be significantly larger than the mean PREP score for the lookback groups.

3. The mean PI score for the embedded question groups will be significantly smaller than the mean PI score for the no-embedded question groups.

4. The mean PI score for the lookback groups will be significantly smaller than the mean PI score for the lookback groups.

5. The mean criterion test score for the embedded question groups will be significantly larger than the mean criterion test score for the no-embedded question groups.

6. The mean criterion test score for the lookback groups will be significantly larger than the mean criterion test score for the lookback groups.

In the PREP probe readers are asked to make both categorical judgments (yes/no rereading is necessary) and confidence decisions (their confidence that their categorical decision was correct). When, as in this experiment, calibration is defined as the correlation between subjective judgments of performance and actual performance, then calibration is most appropriately a measure of the correspondence between the categorical decisions and objective performance; confidence judgments are an auxiliary consideration.

As part of the supplementary analysis, a dependent measure of calibration was created using only the categorical portion of the variable, and results from this measure were compared with results from the two hypotheses using the PREP measure. The influence of these two sources of variance (calibration and confidence) could then be evaluated independently. As a predictor of calibration, it was anticipated that categorical decisions alone would be as accurate as categorical decisions and confidence estimates combined.

Design

The experiment was prepared as a two-factor design consisting of embedded questions and text reinspection. The purpose of the lookback factor was to separate the effects of embedded questions on perceptions of cognitive readiness when combined with re-study decisions from the effects of embedded questions when re-study was prohibited. The factors were crossed, producing four experimental conditions--embedded questions with lookbacks allowed (EQLB), embedded questions with no-lookbacks (EQLB), no-

embedded questions with lookbacks allowed (NEQLB), and no embedded questions with no-lookbacks (NEQNLB). Each factor was between subjects.

Results

The primary question of interest in this experiment was whether embedded questions would improve prediction measures of calibration. The perceived need for strategic remedial behavior was measured using the PREP instrument, which, when compared with objective performance, formulates the dependent variable in the first two research hypotheses. Although there was some difference between embedded question conditions as a function of the lookback option, this two-way interaction was not significant, $F(1,164) = .04$, $p > .05$. Because there were no significant interactions in any of the hypotheses tested, only tests of main effects are appropriate. The variance of the errors of all values of the predictor variable for each of the three dependent variables (PREP, PI, criterion test) appears constant. For this reason, analysis of variance (ANOVA) was used instead of a Brown-Forsythe ANOVA in each test of main effects.

For hypothesis 1, the mean PREP score was significantly greater for the EQ groups (9.25) than for the NEQ groups (6.25), a mean difference of 3.00, $F(1,164) = 12.3$, $p < .05$. For hypothesis 2, there was no significant difference between the mean PREP score for the LB groups (7.56) and the NLB groups (7.94), a mean difference of .38, $F(1,164) = .20$, $p > .05$.

Insert Table 1 about here

Two-way ANOVA was used in hypotheses 3 and 4, where the dependent variable was prediction inaccuracy score (PI). For hypothesis 3, the mean PI score was significantly smaller, and hence more accurate, for the EQ groups (-17.87) than for the NEQ groups (-26.48), a mean difference of 8.61, $F(1,164) = 7.70$, $p < .05$. For hypothesis 4, there was no significant difference between the mean PI score for the LB groups (-21.63) and the the NLB groups (-22.72), a mean difference of 1.10, $F(1,164) = .72$, $p > .05$.

Insert Table 2 about here

The purpose of the last two hypotheses was to test the impact of embedded questions on criterion examination. For hypothesis 5, the mean test score was significantly greater for the EQ groups (51.28) than for the NEQ groups (44.54), a mean difference of 6.75, $F(1,164) = 5.22$, $p < .05$. For hypothesis 2, there was no significant difference between the mean test score for the LB groups (50.35) and the NLB groups (45.47), a mean difference of 4.88, $F(1,164) = 2.73$, $p > .05$. Kuder Richardson 20 was used to calculate reliability for this 17-item test ($r = .56$). Item difficulties ranged from relatively easy (84% correct) to relatively difficult (23% correct), and the average item difficulty was 48% correct with a standard deviation of .18.

Insert Table 3 about here

Subjective beliefs and judgments regarding cognitive readiness were more accurate, better calibrated, for subjects who encountered embedded questions than for those who did not. A main effect for embedded questions was found with both the PI and PREP measures of prediction calibration. In addition, subjects who encountered embedded

questions had higher test scores on the criterion examination than those who did not. No main effects were associated with the look back factor for the three dependent measures.

Predicted Correct. The PI variable, as a measure of prediction calibration, was derived by taking the difference between the scores readers predicted they would receive on the criterion test (subjective judgements) and the scores they obtained (objective measures).

Insert Table 4 about here

One feature immediately discernable when comparing subjective judgments (Table 4) and objective performance (Table 3) is the is the overconfidence displayed by all groups, who, on average, overestimated how well they would do on the criterion examination by 22%, a clear example of poor calibration. Not all subjects overestimated how well they would do on the examination however; whereas 18 subjects in the EQ conditions scored better on the examination than they had predicted, only 7 subjects in the NEQ conditions made such underestimations.

Two measures of prediction calibration were used in this study (PI and PREP). The first instruction subjects received after reading the text was to estimate the score they would obtain on the examination (PI). Immediately after responding, they were asked about their perceived need to reread at the 20, 40, 60, and 80% criterion level (PREP). Although encountering embedded questions produced significant main effects with both measures, the effects were stronger when PREP was the independent variable ($d = .57$) than when it was PI ($d = .43$). A standardized mean difference, $\underline{d}$, was used to compare effect sizes because the standard deviations between the two variables differed greatly and a natural common scale is interpretable for the PI measure but not for the PREP measure (Green & Hall, 1984).

Decisions associated with the PREP variable were generally more conservative than, and often in contradiction with, decisions made in conjunction with the PI measure. For example, 49 subjects predicted test scores of 81% or higher, but when asked on the PREP measure if they thought they would need to reread to obtain a score of 80%, 16 of the 49 subjects said they would. At the 60% criterion level, 124 subjects predicted test scores of 61% or higher, but when asked on the PREP measure if they thought they would need to reread to obtain a score of 60%, 32 of the 124 subjects indicated that they would. In the same direction, 28 subjects contradicted themselves at the 40% probe and 17 at the 20% probe. Because poor calibration is generally the result of overconfidence, the more conservative decisions associated with the PREP measure are the likely source of larger effects.

Reasons why more conservative decisions were made in connection with the PREP measure are speculative, of course. Although PI and PREP are conceptually affiliated, psychologically they ask subjects to make somewhat different decisions; in the PI probe subjects were asked to make a subjective estimate of what they thought they could do on the test, whereas the PREP probes asked them to estimate whether strategic behavior (rereading) would be necessary to make sure they could reach four specific criterion levels. The subjective estimates made in association with the PI measure, although reflecting here-and-now perceptions of cognitive readiness, were probably also affected by the reader's past performance--"I am the type of person who generally scores __% on examinations." According to MacKenzie (1989), learners' best estimate of performance, in the absence of other stronger cues, will be their mean past performance.

The probes associated with PREP and PI obviously tapped different sources of subjective feelings, and this distinction served to mitigate the empirical relationship between the two variables. The correlation between the two variables was statistically significant and in the predicted direction; however, the strength of the relationship was modest

(r = .38). The possible explanation that PREP is both a measure of calibration and confidence, whereas PI is a measure of calibration only, is insupportable, however, given that the correlation between PI and the calibration portion of PREP alone was r = .33.

As mentioned earlier, an analysis was undertaken of the categorical decisions associated with PREP alone. In the PREP measure, readers were asked if they would need to reread if the criteria for passing the examination were 20, 40, 60, and 80%. Four correct decisions are possible.

Insert Table 5 about here

By comparing Table 1, where PREP was analyzed using both the calibration and confidence components of its measure, with Table 5, where the calibration (categorical) decisions are considered alone, it is possible to discern that main effects for the embedded question conditions are nearly identical regardless of approach. Standardized main effects for the embedded questions conditions in the two component method was d = .57; with calibration decisions considered alone, the effect size was d = .50. In this experiment, embedded questions improved accuracy of decision making but had little impact on how readers use the confidence scale. Moreover, differences in effect sizes between the PI and PREP variable had little to do with the confidence component of the PREP measure.

Strategic Processing Variables. Data were also collected on variables not part of the formal research hypotheses: The amount of time subjects took to take the Xenograde examination, the amount of time they spent reading the text, and the number of times readers turned the pages backwards in the two lookback conditions. On average, subjects spent 17.10 minutes completing the 17 items in the Xenograde examination, with a standard deviation of 5.26 minutes. No interaction or main effects were present.

In comparison, subjects spent an average of 15.22 minutes reading and studying the Xenograde text. Spending less time preparing for an examination than actually taking it is probably indicative of both the amount of interest subjects had in learning the experimental materials and overall poor calibration of test readiness. The two-way interaction between embedded question conditions as a function of the lookback option was not significant, but reading times were significantly different in both in embedded questions and lookback tests of main effects.

Insert Table 6 about here

The two EQ groups were engaged in reading the Xenograde material an average of 4.04 minutes, or 31%, longer than the two NEQ conditions. However, the embedded questions added 280 words to the 1700 word Xenograde text, a 16% increase in the amount of text to be read. Because the EQ groups spent 31% more time reading 16% more text, embedded questions probably altered reader perceptions regarding the minimum amount of cognitive effort needed to comprehend the Xenograde material.

Readers permitted to look back averaged 1.21 more minutes reading than the NLB groups, and as reported in Table 6, this difference was statistically significant. When pairwise tests were conducted comparing the EQLB group with the EQNLB group ($F(1,82) = 6.86$, $p < .05$) and the NEQLB group with the NEQNLB group ($F(1,82) = .25$, $p > .05$) it was possible to discern that the source of this main effect was with the first comparison. That is to say, having the option to look back changed the amount of time on task only when embedded questions were present. Apparently, embedded questions changed reader perceptions of the minimum amount of cognitive effort needed to understand the experimental material. And, when given the opportunity to remediate their understanding through the lookback option, they did.

Discussion

The central question of this study was whether embedded questions could bring subjective judgments of test readiness in closer correspondence with objective assessment of test performance--prediction calibration. As hypothesized, embedded questions improved prediction calibration judgments of calibration. These effects were discovered with two different, but conceptually related, measures (PI and PREP). With the PI measure, readers who encountered embedded questions gave more accurate evaluations of how well they would do on the criterion test. With the PREP measure, readers who encountered embedded questions had more accurate perceptions regarding how much strategic behavior (rereading) would be necessary to reach four different levels of performance. The source of the EQ main effects on PREP were found with the categorical decisions and not with the confidence portion of the measure. In addition, readers who encountered the embedded questions performed significantly better on the criterion examination than those who did not encounter the questions.

Although readers in the NEQLB group used the lookback option, they did not spend any more time on task than the NEQNLB group. The EQ groups spent 31% more time reading 16% more text when compared with the NEQ groups. Clearly, embedded questions helped readers realize that greater effort was required. Moreover, readers who both encountered embedded questions and were given the option to remediate their understanding through the lookback option did lookback and ultimately spent more time engaged in studying the experimental materials. Given that research findings consistently show that readers generally overestimate their sense of preparedness, these findings are noteworthy and encouraging.

Our results are especially meaningful in the face of Maki and Serra's (1992) findings that practice tests similar or identical to criterion tests did not improve readers' prediction accuracy. Two reasons appear plausible. First, embedded questions engage the

reader during the task of reading and are directly related to the text itself, whereas practice tests are taken after the fact and may be perceived by the reader as either a reliable or unreliable guide to the final criterion measure. Embedded questions are written such that their relevance is clearly evident; practice tests require interpretations as to their ultimate relevance. Second, Maki and Serra utilized multiple choice questions, whereas Davey (1987), Ghatala et al. (1989), and Pressley et al. (1990) found that practice tests consisting of short answer questions produced stronger levels of calibration than tests consisting of multiple choice questions. The embedded questions used in our experiment, and the type we suggest for inclusion in text, required only short answers.

Generalizing from this study to nonexperimental conditions seems warranted for two reasons. First, subjects were not cued to the presence of embedded questions nor were they asked to answer them. Because reading for remembering is a learner-controlled process, and because the effects of embedded questions ultimately depends on what the reader does with them, we sought to determine if embedded questions could induce a spontaneous, learner-produced versus investigator-induced effect. This design feature makes the study particularly unique and also allows for the greatest degree of generalizability to nonexperimental settings. Second, the facilitating effects of embedded questions were found in an experiment of very short duration, with subjects whose only motivation was probably to get through the experiment. It is reasonable to speculate that the effects of embedded questions may be more robust under conditions more similar to the types of academic reading, testing, and incentive situations experienced by adults.

Implications and Directions for Future Research

The evidence is compelling that many readers get through academic courses without acquiring a clear understanding of the most fundamental aspects of the material the text is intended to communicate. The most serious problem is not so much readers' inherent inability to read, but rather their interaction with the text. Embedded questions have a rich

history of assisting learners in acquiring concepts and principles from prose passages. What our findings suggest is that embedded questions can be used to change the dysfunctional interaction many readers have with the text. Well thought out embedded questions have the potential to challenge readers' understanding of what they are reading while they are engaged in the process of reading. This is in marked contrast to current practices where the readers' first test of their understanding is at the time of formal testing. Readers of all ages have a repertoire of strategies they can employ to remediate their understanding. However, they have no reason to use them if they do not understand that they do not understand.

Our research findings support the conclusion that embedded questions have the effect of bringing subjective beliefs regarding test readiness in better calibration with objective test preparedness. Being well calibrated has powerful advantages; however, meaningful learning is cumulative, and the ability to learn new material is highly dependent upon prior knowledge. For this reason, small differences in cognitive monitoring ability may account for large differences in academic performance if considered over the course of several school years. Regardless of the causes of poor cognitive monitoring, the consequences are the same: poorly calibrated students will be the least likely to engage themselves when the academic situation demands it most. After a short duration, the differences between good and poor calibrators will expand, and those with poor cognitive monitoring skills will also have an impoverished prior knowledge upon which to learn new material. In this experiment, embedded questions positively altered reader perceptions of cognitive readiness in a reading setting that lasted less than a half hour. When considering their effects over a longer duration, embedded questions can serve to mitigate the cumulative harm that results from poor cognitive monitoring.

If calibration plays an important role in the process of reading and understanding, it is logically imperative that classroom teachers teach for calibration. Readers must be taught

to self-question and self-cue to bring forth information relevant to metacognitive control. Embedded questions can be though of a as a prompt whereby students ask themselves questions as a test of their understanding of the text. If used often enough, and under conditions were there is a legitimate connection between successfully answering the embedded questions and doing well on the criterion test, they will help students internalize self questioning and take more control of their own learning.

After noting how subjects interacted with the material in this experiment, we believe that cognitive tempo may be an individual difference informative to future research. Some readers may be poorly calibrated because they lack deliberateness in testing their understanding (Kagan & Koran, 1970); others might be more accurately described as defensive and anxious and choose to escape the stressful act of evaluating their understanding by making quick decisions about their state of cognitive readiness (Wapner & Conner, 1986). With a longer investigation, measures of individual differences could be tested, as could the relationship between aptitudes and performance at different stages of learning. Ultimately, techniques and measures of on-line cognitive processing are required--cognitive measures taken at the moment of learning.

In this experiment embedded questions positively altered perceptions of cognitive readiness and had the effect of making readers better calibrated. The logical next step is to determine how these effects are produced. Embedded questions may provide feedback that readers can use to make adjustments in their judgments of cognitive readiness. They may also act as a prosthetic device, triggering readers to evaluate their comprehension where they may not have done so on their own. However, the important point is that success in isolating any of these cognitive processes will depend on more powerful research designs, more sensitive measures, and data collection taken from real-world academic settings. Semester-length research would allow for both more stable measurements and more powerful within-group designs.

REFERENCES

Andre, T. (1987). Questions and learning from reading. Questioning Exchange, 1(1), 47-86.

Baker, L. & Brown A. (1984a). Meta-cognitive skills in reading. In P. D. Peterson (Ed.), Handbook of reading research (pp. 353-395). New York: Longman.

Baker, L. & Brown A. (1984b). Cognitive monitoring in reading. In J. Flood (Ed.), Understanding reading comprehension: Cognition, language, and the structure of prose, (pp. 21-44). Newark, DE: International Reading Association.

Brown, A. L., Brandsford, J., Ferrara, R., & Campione, J. (1982). Learning, remembering, and understanding (Report No. 224). Urbana-Champaign: University of Illinois Center for the Study of Reading.

Brown, A., Armbruster, B., & Baker, L. (1986). The role of metacognition in reading and studying. In J. Orasanu (Ed.), Reading comprehension: From research to practice (pp. 49-75). Hillsdale, NJ: Erlbaum.

Campbell, D. T. & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Skokie, IL.: Rand McNally.

Cronbach, L. & Snow, R. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

Epstein, W., Glenberg, A., & Bradley, M. (1984) Coactivation and comprehension: Contributions of text variables to the illusion of knowning. Memory & Cognition, 12, 355-360.

Glenberg, A. & Epstein, W. (1985). Calibration of Comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 702-718.

Glenberg, A. & Epstein, W. (1987). Inexpert calibration of comprehension. Memory & Cognition, 15, 84-93.

Glenberg, A., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. Journal of Experimental Psychology: General, 116, 119-136.

Glenberg, A., Wilkinson, A., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. Memory and Cognition, 10, 597-602.

Green, B. F. & Hall, J. A. (1984). Quantitative methods for literature reviews. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology (Vol 35. pp. 37-53). Palo Alto, CA: Annual Reviews.

Hare, V. C. & Smith, D. C. (1982). Reading to remember: Studies of metacognitive reading skills in elementary school-aged children. Journal of Educational Research, 75, 157-164.

Kagen, J., & Koran, N. (1970). Individual variation in cognitive processes. In P. Mussen (Ed.), Carmichael's manual of child psychology (3rd ed., Vol 1). New York: Wiley.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art. In H. Jungerman & G. de Zeeuw (Eds.), Decision making and change in human affairs (pp. 1-13). Amsterdam: Reidel.

MacKenzie, R. (1989). Thinking and judgment: What they are and how to teach them. Gainesville, FL: RSM Press.

Maki, R. & Berry, S. L. (1984). Metacomprehension of text material. Journal of Experimental Psychology: Learning Memory and Cognition, 10, 663-679.

Maki, R. & Serra M. (1992). Role of practice tests in the accuracy of test predictions on text material. Journal of Educational Psychology, 84, 200-210.

Merrill, M. D. (1965). Correction and review on successive parts in learning a hierarchical task. Journal of Educational Psychology, 56, 225-234.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 288-294.

Pratt, M. W., Luszcz, M., MacKenzie-Keating, S. & Manning, A. (1982). Thinking about stories: The story schema in metacognition. Journal of Verbal Learning and Verbal Behavior, 21, 493-505.

Pressley, M., Snyder, B., Levin, J., Murray, H., & Ghatala, E. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. Reading Research Quarterly, 22, 219-236.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. Journal of Verbal Learning and Verbal Behavior, 20, 216-230.

Wagoner, S. (1983). Comprehension monotoring: What it is and what we know about it. Reading Research Quarterly, 18, 328-346.

Walczyk, J. & Hall, V. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. Journal of Educational Psychology, 81, 435-437.

Wapner, J., & Conner, K. (1986). The role of defensiveness in cognitive impulsivity. Child Development, 57, 1370-1374.

Zimmerman, J., Broder, P., Shaughnessy, J., & Underwood, B. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. Intelligence, 1, 5-31.

Figure 1

Probe	Reader Decision Would/would not need to reread	Correct Decision	Confidence Score	How Scored
20%	Would not	Yes	5	+5
40%	Would not	Yes	2	+2
60%	Would	No	2	-2
80%	Would	Yes	4	+4

Table 1.

Mean PREP Score by Experimental Condition

Group	N	M	SD
EQLB	42	9.14	4.42
EQNLB	42	9.36	5.38
NEQLB	42	5.98	6.52
NEQNLB	42	6.52	5.67

Two-way ANOVA for PREP Scores by Experimental Group

Source	df	ss	ms	F
EQ (A)	1	378.00	378.00	12.29*
LB (B)	1	6.10	6.10	.66
AB	1	1.17	1.17	.85
Error	164	5044.24		

* $p < .05$.

Table 2.

Mean PI Score by Experimental Condition

Group	N	M	SD
EQLB	42	-17.91	18.46
EQNLB	42	-17.84	21.89
NEQLB	42	-25.35	22.64
NEQNLB	42	-27.61	16.85

Two-way ANOVA for PI Scores by Experimental Group

Source	df	ss	ms	F
EQ (A)	1	3110.62	3110.62	7.70*
LB (B)	1	50.71	50.71	.72
AB	1	57.05	57.05	.71
Error	164	66273.20	404.11	

* $p < .05$.

Table 3.

Mean Examination Scores by Experimental Condition

Group	N	M	SD
EQLB	42	54.91	20.58
EQNLB	42	47.66	18.31
NEQLB	42	45.80	20.46
NEQNLB	42	43.27	17.04

Two-way ANOVA for Examination Scores by Experimental Group

Source	df	ss	ms	F
EQ (A)	1	1912.95	1912.95	5.22*
LB (B)	1	1002.06	1002.06	.10
AB	1	233.59	233.59	.64
Error	164	60163.49		

* $p < .05$.

Table 4.

Mean Predicted Scores by Experimental Condition

Group	N	M	SD
EQLB	42	72.81	19.41
EQNLB	42	65.5	21.08
NEQLB	42	71.14	19.12
NEQNLB	42	70.88	15.83

Table 5.

Mean Number of Correct PREP Decisions by Experimental Condition

Group	N	M	SD
EQLB	42	3.14	.61
EQNLB	42	3.17	.82
NEQLB	42	2.74	.83
NEQNLB	42	2.79	.84

Two-way ANOVA for Number of Correct PREP Decisions by Experimental Condition

Source	df	ss	ms	F
EQ (A)	1	6.48	6.48	10.61*
LB (B)	1	.05	.05	.09
AB	1	.01	.01	.01
Error	164	100.17		

* $p < .05$.

Table 6.

Average Minutes Reading the Text by Experimental Condition

Group	N	M	SD
EQLB	42	18.26	3.46
EQNLB	42	16.24	3.62
NEQLB	42	13.41	4.16
NEQNLB	42	13.02	2.73

Two-way ANOVA for Average Minutes Reading the Text by Experimental Condition

Source	df	ss	ms	F
EQ (A)	1	684.05	684.05	54.97*
LB (B)	1	60.72	60.72	4.88*
AB	1	28.34	28.34	2.28
Error	164	2040.83		

* $p < .05$.